A Predictive Analytics Approach to Identify "Best Value" Graduate Programs in Analytics & Data Science

Gavin Ellis, Christopher Embry, Griffin Koester, Josh Swan, Michael Will, Matthew A. Lanham

Purdue University, Krannert School of Management

ellis103@purdue.edu; embryc@purdue.edu; gkoester@purdue.edu; swanj@purdue.edu; will7@purdue.edu; lanhamm@purdue.edu

ABSTRACT

This study uses predictive analytics to identify "best value" graduate analytics and data science programs. The motivation for this study is that many of these types of programs are often very expensive, averaging close to $40,000 with some costing over $100,000. We define best value as the average yearly starting salary minus the cost of the program. Data was collected from 197 graduate analytics, business analytics, and data science program websites. We then developed a predictive model to predict starting salaries of graduates from a program using features about the program curriculum, experiential opportunities, student demographics, and other factors. From our model we identify which programs are "best value" graduate programs by accounting for program cost and length. We posit our analysis will be of particular interest to students considering pursuing these types of degrees, as well as higher education administrators seeing where their programs stand in the market.

INTRODUCTION

Nearly 250 graduate programs now exist for data analytics (1). Given that many of these programs are relatively young, frameworks need to be created to help schools and applicants understand the connection between program features and labor market outcomes. This project focuses on a value proposition, seeking to predict how factors including program costs and time commitments relate to job placements and starting salaries.

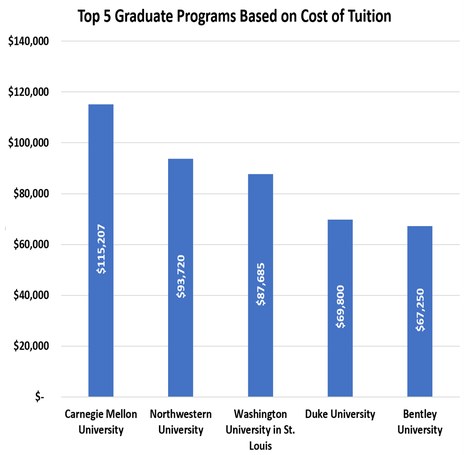

Fig 1. Graduate School Cost

Research Questions:

- How do graduate analytics program features drive eventual career outcomes and value?
- Which graduate analytics programs have the best ROI?

1. https://analytics.ncsu.edu/?page_id=4184

LITERATURE REVIEW

Our study combines an evaluation of multiple graduate program factors, predictive modeling methods, and a cost-benefit analysis to build a novel decision-making tool for potential graduate students in the analytics and data science fields. Existing studies employ similar techniques, but for broader fields, and lack tools for graduate students to decide if a program is right for them, or if it is worth the investment.

Aspects of Our Study:	Valletta (2015)	English, Umbach (2016)	Ma, Pender, Welch (2019)	Abel, Deitz (2014)	Tran (2017)
Grad School / BA Focus	✓	✓			✓
Multi-Factor Evaluation	✓	✓	✓	✓	✓
Predictive Modeling		✓		✓	
Decision-Making Tool				✓	
Cost / Benefit Analysis		✓	✓	✓	✓

METHODOLOGY

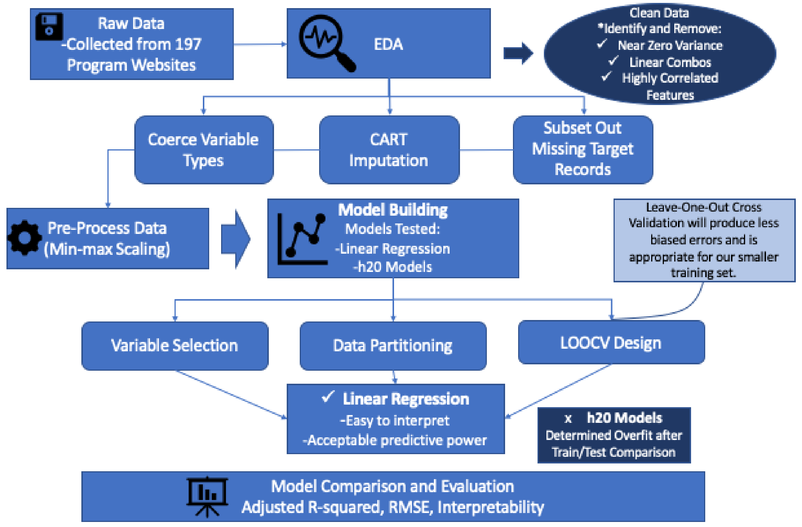

Fig 2. Methodology Flowchart

STATISTICAL RESULTS

Our team investigated several statistical and machine learning algorithms. Upon comparing the RMSE and Adjusted R2 for these models, it became apparent that the more flexible models had a greater accuracy on the training data but were severely overfit when the model was applied to the holdout set. Despite being less flexible, the multiple regression model displayed the most consistent results across both sets of data. We found that the largest drivers of future salary are what type of school the degree is from, the student's GPA, and the availability of certain courses. Figure 3 displays a series of charts indicating that the assumptions of OLS regression have been met.

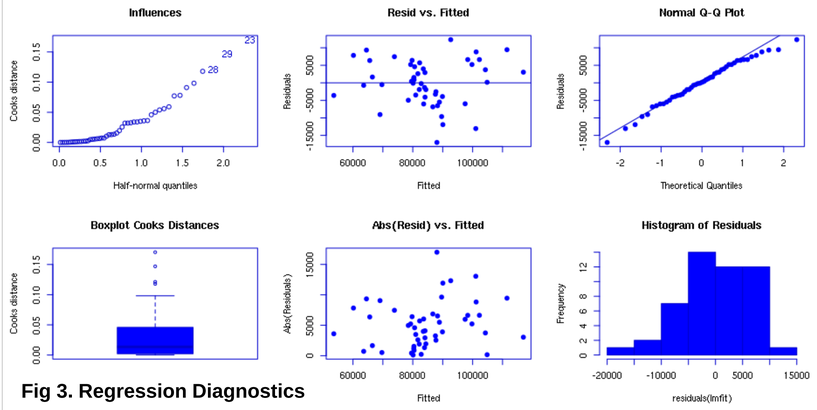

Fig 3. Regression Diagnostics

BUSINESS IMPACT

Our model provides prospective graduate students with an understanding of what program factors are most significant to predicting their future salary. By knowing their possible starting salary upon graduation, students will be able to maximize their return on investment. Additionally, we believe these findings can add value to academic administrators. Our final model for predicting starting salary has an adjusted R-squared value of 0.8075, which can allow program directors to reliably develop programs that will lead to student success. Figure 5 indicates that graduate schools should invest their money on implementing **Cloud AWS** classes, **Python programming**, and opportunities to complete **industry projects** because they have the largest impact on projected salary. These findings should enable universities to provide programs that will best equip their students for a career in the industry.

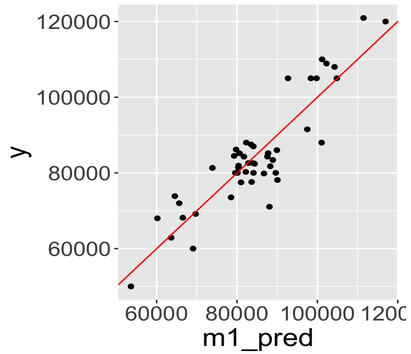

Fig 4. Predicted Starting Salary Plot

Significant Factors	Effect on Predicted Salary
GPA	6526
Cloud, AWS Classes	6255
Python	5661
Industry Projects	5415
Placement Rate	-3390
Science School	-3981
Text Analytics, NLP Classes	-4372
Webmining	-4985
STEM Certified	-5521
Business School	-6126
Predictive Analytics Classes	-7747

Fig 5. Significant Factors

CONCLUSIONS

- Several factors such as student GPA, courses in cloud computing, and school type are all significant predictors at the 95% confidence level.
- Our model can be expanded upon to help prospective graduate students calculate the return on investment for attending different programs.
- It must be noted that our model was significantly limited by the availability of data.
- Out of 197 schools, our team was only able to utilize a sample size of 53.
- These limitations severely impacted the accuracy of our model, and further data collection is needed to confirm these results.

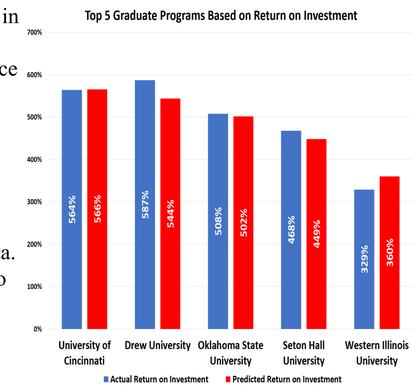

Fig 6. Predicted and Actual ROI

ACKNOWLEDGEMENTS

We would like to thank Professor Matthew Lanham, Xinyu Wang, Theo Ginting, and our graduate student mentors Chou Sheng-Yang, Cao Niuying, and Ren Ruihua for their guidance and support on this project.